



Early Journal Content on JSTOR, Free to Anyone in the World

This article is one of nearly 500,000 scholarly works digitized and made freely available to everyone in the world by JSTOR.

Known as the Early Journal Content, this set of works include research articles, news, letters, and other writings published in more than 200 of the oldest leading academic journals. The works date from the mid-seventeenth to the early twentieth centuries.

We encourage people to read and share the Early Journal Content openly and to tell others that this resource exists. People may post this content online or redistribute in any way for non-commercial purposes.

Read more about Early Journal Content at <http://about.jstor.org/participate-jstor/individuals/early-journal-content>.

JSTOR is a digital library of academic journals, books, and primary source objects. JSTOR helps people discover, use, and build upon a wide range of content through a powerful research and teaching platform, and preserves this content for future generations. JSTOR is part of ITHAKA, a not-for-profit organization that also includes Ithaka S+R and Portico. For more information about JSTOR, please contact support@jstor.org.

A FORMULA FOR DRAWING TWO CORRELATED CURVES SO AS TO MAKE THE RESEMBLANCE AS CLOSE AS POSSIBLE.

BY FRANCIS TODD H'DOUBLER.

One of the problems with which statisticians are confronted almost daily is that of drawing two curves in such a way that the correspondence, so far as the eye can judge, shall be as close as possible. The statistician, as a rule simply experiments with his scales, varying the two units until the curves fall into some sort of correspondence, abandoning his graphical illustration if the appearance of the two curves is not such as to warrant a positive conclusion. As is shown below, it is possible to derive a formula which will, without imposing too much labor upon the statistician, make this resemblance as close as could be desired. If, for instance, we are comparing yearly marriage rates with yearly prices of wheat, the two groups will be brought into closest correspondence if, after plotting the marriage rate, we multiply each annual price of wheat by the expression $\frac{\sum x_i y_i}{\sum y_i^2}$, where $x_1, x_2 \dots x_N$ represent the annual marriage rates, and $y_1, y_2 \dots y_N$ represent the annual prices of wheat. By closest correspondence, in the above statement, is meant such juxtaposition of the two curves as will make the sum of the squares of the distance between the two curves at the successive years a minimum. This definition of closest correspondence has much to be said in its favor on general mathematical grounds. If, however, we define closest correspondence as such juxtaposition as will make the sum of the distances between the two curves regardless of sign, a minimum, the problem becomes much more complex and the solution very laborious. In the great majority of instances, moreover, the adjustment of the two curves by the two methods would probably be practically identical. The problem has possibly been worked out before, but has never come to my attention; and the utility of the formula in practical statistical work warrants its reproduction, even if it has been previously published.

Let $x_1, x_2, x_3 \dots x_N$

and $y_1, y_2, y_3 \dots y_N$

be the statistics of two correlated phenomena, where x_i and y_i correspond to the same value of the variable. Thus x_1 and y_1 might be respectively the marriage rates and the average wheat prices of a certain year; x_2 and y_2 the rates of the next year, and so on.

If these figures are plotted, two curves are obtained which show in a way the correlation existing between the two phenomena. We wish to make the two diagrams fit as closely as possible.

Plot the x 's just as they occur. Then multiply each y by the quantity e , and plot the resulting figures. We have the two sets

$$\begin{array}{l} x_1, x_2, x_3 \dots x_N \\ ey_1, ey_2, ey_3 \dots ey_N \end{array}$$

Consider the sum of the squares of the differences between corresponding x 's and y 's. This quantity is

$$(x_1 - ey_1)^2 + (x_2 - ey_2)^2 + \dots + (x_N - ey_N)^2$$

or, in short

$$\sum_1^N (x_i - ey_i)^2$$

It must be remembered that x_i and y_i are fixed numbers. Evidently

$$\sum_1^N (x_i - ey_i)^2 = f(e)$$

is a function of e —, that is, it varies with e . Let us seek that value of e which causes this function to take its minimum value.

First of all we note that $\sum (x_i - ey_i)^2$ is the sum of squares and hence of positive quantities. Consequently it is always positive. Moreover, it increases beyond limit as e increases. Therefore, when the derivative of this function is equated to zero we have the condition for a minimum (by differential calculus).

Differentiating and equating to zero, we have

$$-2(x_1 - ey_1)y_1 - 2(x_2 - ey_2)y_2 - \dots - 2(x_N - ey_N)y_N = 0$$

and solving for e

$$e = \frac{x_1y_1 + x_2y_2 + x_3y_3 + \dots + x_Ny_N}{y_1^2 + y_2^2 + \dots + y_N^2}$$

or

$$e = \frac{\sum_1^N x_i y_i}{\sum_1^N y_i^2} \quad (1)$$

When the y 's are multiplied by this value of e the sum of the squares of the differences $(x_i - ey_i)$ is a minimum, and the x — curve and the modified y — curve are sure to fit closely.

In case the y 's are unaltered while the x 's are all multiplied by e^1 , the value of e^1 giving a minimum to $\sum (y_i - e^1x_i)^2$ is similarly found to be

$$e^1 = \frac{\sum_1^N x_i y_i}{\sum_1^N x_i^2} \quad (2)$$

If $x_i = y_i$, then both e and e^1 are equal to unity.

One might ask why it would not be better to consider merely the differences $(x_i - ey_i)$ regardless of sign, and to seek the value of e causing the sum of these differences to assume its minimum. In other words, consider the sum of the absolute values,* $|x_i - ey_i|$, — *i.e.*

$$|x_1 - ey_1| + |x_2 - ey_2| + |x_3 - ey_3| + \dots + |x_N - ey_N|$$

or

$$\sum_{i=1}^N |x_i - ey_i|$$

This function is always positive. Its graph is a broken line with N corners—a corner being located above each of the abscissae, $e = x_i / y_i$ ($i=1 \dots N$). One of these N ratios gives

$$\sum_{i=1}^N x_i - ey_i, \text{ its minimum. It must be found by actually}$$

substituting each ratio and seeing which gives the above sum the smallest value. There is no general formula applicable here, since the derivatives at these corners are all discontinuous. This method involves so much tedious work that the first method, with its general formula, is much more practicable, as well as preferable.

*In mathematical notation $|x|$ is used to indicate the absolute or numerical value of x . Thus $|-10| = |10|$.